COVID-19 Analysis

Maeva Braeckevelt

Executive summary

This analysis aimed at analyzing the pattern of association between registered COVID-19 cases and registered death due to COVID-19 until the 26 October 2020 among countries. The data used was gathered both by the Center for Systems Science and Engineering and the World Bank'sites. The main variables that I used were: total of registered number (per thousand people), total of death (per thousand people) and the population (per millions). The regression model chosen was the Log-Log Weighted Linear regression. It showed a linear pattern between death and confirmed case, such as, for people among all countries, for +10% change in registered COVID-19 case, there is an association of +9,5% change in registered Covid-19 deaths. However, this analysis is subject to the the politics testing of every country (mass testing, few tests, etc.)

Introduction

The aim of this project is to analyze the pattern of association between the confirmed case of Covid-19 and the death du to it. The variables I will use in this analysis are the confirmed case per thousand people (x), deaths per thousand people (y) and the population in millions of 170 countries until the 26/10/2020. The population is all the covid-19 confirmed case and all the death due to it. There are 195 countries in the world, so my sample of 170 is representative. However, the quality of my data depends of the accuracy of the counting done by every country. We can't be sure that every country had the same counting process.

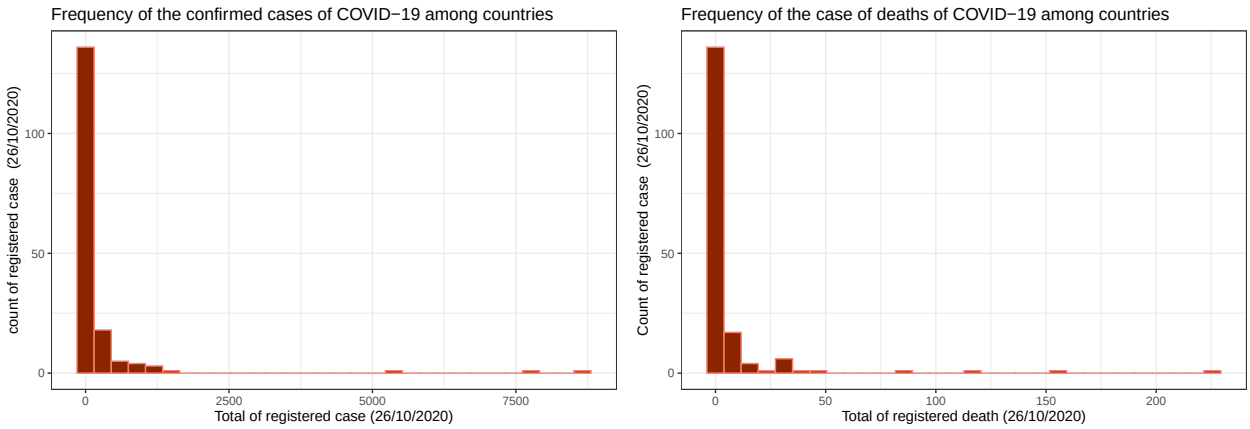
Observations

To simplify the understanding of my analysis, I have decided to drop the countries that have 0 death. It only represent 7% of the data so I will focus on the majority of non-zero values. I decided to convert my cases confirmed and of deaths per 1000 people, and my population by millions. These conversion will help the clarity of the analysis and give me the possibility of using the log-transformation if needed.

Histogram and summary statitics

Table 1: Summary statistics

variable	mean	median	min	max	sd	skew
Case confirmed	252.983677	27.978	0.033	8660.088	998.89399	7.041386
case of death	6.784735	0.431	0.001	225.458	24.59623	6.376255



The summary statistic does not take to account the size of the population of the countries, so the interpretation is less meaningful. There is very small and very big population. But I observed some similarities between the two variables : they both are skewed with a right tail and some extreme values. The median of both are significantly smaller than the mean. The majority of the countries has low value. Although those variable share the same tendency, the scale is very different between them. I can already sense that I will have to find a way to weight the population to be able to have a relevant analysis.

Transformation of the variables

To uncover the trend of the pattern association, I investigated four non-parametric regressions : level-level (figure A1 in the appendix), log-level (A2), level-log (A3), log-log(A4). I chose to use the fourth model (A4) : use Log transformation for both of my variables.

Substantive reasons : the COVID-19 is a very contagious disease, one person can infect multiple people, so the variable are affected in multiplicative ways. The graph itself fits better for the interpretation. Also, we are looking for percentage association.

Statistical reasons: The distributions of the variables are skewed with a long right tail. So taking the log is good solution to make the distribution of my transformed variable more symmetric. In addition,my variables have the same metric so for the purpose of comparison, taking log for both is easier. The graph (A4) can be interpret in a meaningful way and give a good prediction.

Model choice and interpretation

The regression model that I have chosen is the Log-Log Weighted Linear regression : weight by coutries' population. It sounded pertinent that bigger population has a bigger impact on the slope. Please find in the appendix the estimation of the different models and the argumentation.

Formula : ln_death = -3,39 + 0,95 * ln_confirmed, weights: countries'population

Alpha : -3,39 is the average of ln_death when the confirmed case is one (ln(1)=0). **Beta** : the deaths is 9,5 percent higher on average for observation having 10 percent higher case confirmed In log-log transformation, alpha is usually not meaningfull. The deaths is not increasing as fast as the confirmed cases. The graph (A8) shows that bigger the population is bigger are the cases of deaths. It will be very interesting to transform the variable per capita. This way, I could take out the bias having more deaths by higher population.

Hypothesis testing on Beta

Table 2: Hypothesis testing on ß

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	-3.3919426	0.3728914	-9.096329	0	-4.1280992	-2.655786	168
ln_confirmed	0.9498549	0.0610875	15.549077	0	0.8292568	1.070453	168

H0 : there is no pattern association between deaths and confirmed cases.

HA : there is a pattern of association between deaths and confirmed cases.

I chose 95% confidence interval for the hypothesis testing. My confidence interval is not crossing 0, so it means that it is significant. My p value is $< 0{,}05$ so I can reject my HO and therefore there is a pattern of association between deaths and confirmed cases on the 26/10/2020.

$$H_1 : \beta \neq 0$$

Analysis of the residuals

Table 3: List of the 5 countries with the largest negative errors

country	ln_deaths	reg4_y_pred	reg4_res
Burundi	-6.907755	-3.9477882	-2.959967
Liechtenstein	-6.907755	-4.3492614	-2.558494
Qatar	-1.469676	1.2400196	-2.709696
Singapore	-3.575551	0.4643979	-4.039949
Sri Lanka	-4.135167	-1.4320953	-2.703071

For this 5 countries, the model overestimated the deaths. I can see that the predicted value is smaller than the real value. For example, for Liechtenstein , the ln_deaths is equal -6,91, but the predicted value for is -4.35 so the predicted value is off by -2,56. So those countries have less death than the average. There could be a lot of explanation for it : better healthcare, less sensitif to covid19 (young people), isolation, etc.

Table 4: List of 5 countries with the largest positive errors

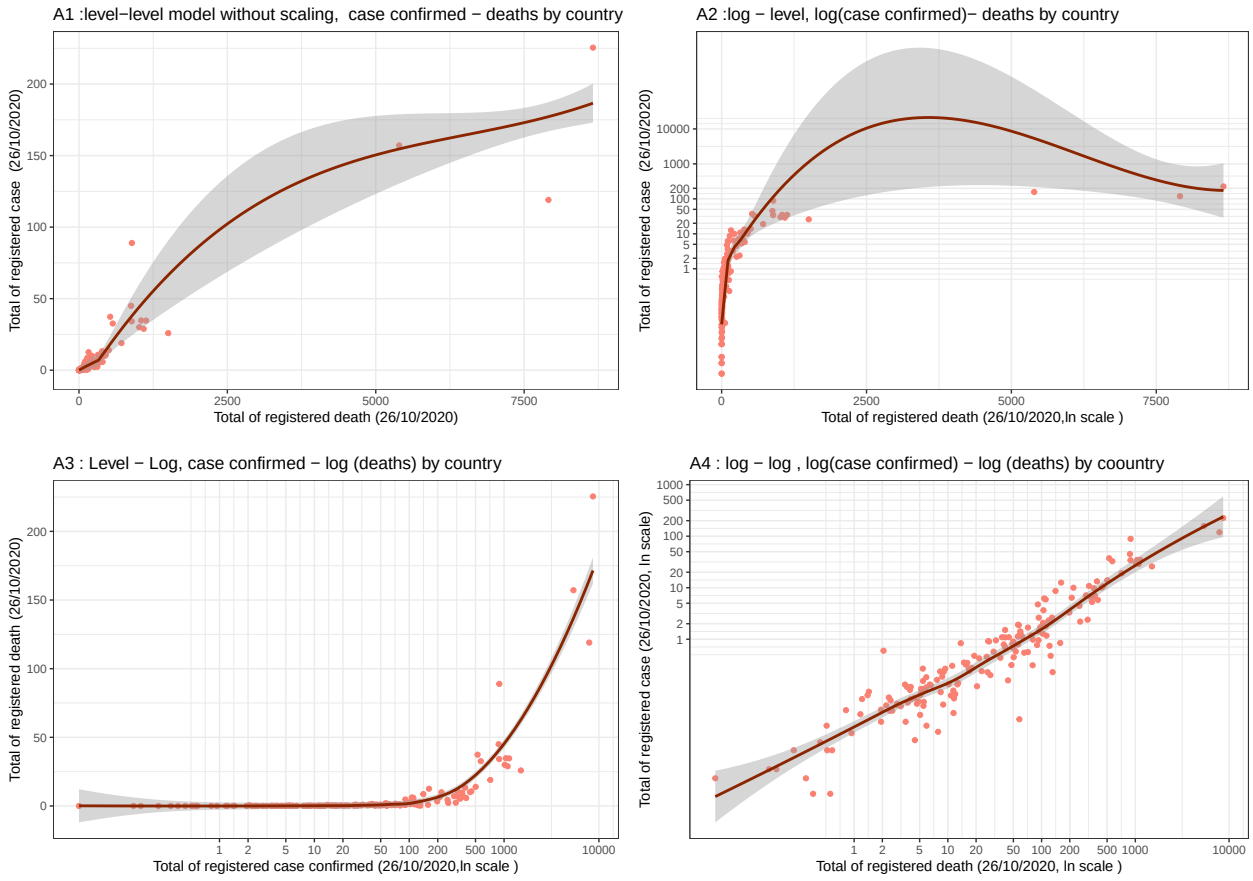
country	ln_deaths	reg4_y_pred	reg4_res
Ecuador	2.5299597	1.438394	1.091566
Iran	3.4848030	2.633650	0.851153
Italy	3.6200116	2.558791	1.061220
Mexico	4.4877821	3.059970	1.427812
Yemen	-0.5124937	-2.705477	2.192983

For this 5 countries, the model underestimated the deaths. I can see that the predicted value is bigger than the real value. For Italy, the ln_deaths is equal to 3,62, the predicted value is 2,55 so the predicted value is off by +1,06. So, Italy has more deaths than the average. It could be due to the way they handle the pandemic, or the bad healthcare, or old population, etc

Appendix

This appendix contains the documentation of the analysis annoted.

Investigation of the transformation of the variable



Estimating different models

I estimated fourth model : Simple linear regression (A5), quadric linear regression (A6), Piecewise linear spline regression (A7) and Weighted linear regression : weight with population (A8).

Simple Linear regression (A5)

This graph represents the simple linear regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -4,12 + 1,03 * ln_confirmed

Alpha : -4,01 is the average of ln_death when the confirmed case is one (ln(1)=0).

Beta : the deaths is 10,3 percent higher on average for observation having 10 percent higher case confirmed

In log-log transformation, alpha is usually not meaningfull. The deaths is increasing faster than the confirmed cases.

The adjusted R squared is 0,89. That's a trustful model.

Quadric Linear regression (A6)

This graph represents the quadric linear regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -4,01 + 0,88 * ln_confirmed + 0,02 * ln_confirmed^2

Quadric Linear Regression is very hard to interpret The deaths is not increasing as fast as the confirmed cases.

The adjusted R squared is 0,89. That's a trustful model

Piecewise linear spline regression (A7)

This graph represents Piecewise linear spline regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020. I chose two cut-off, the first at one, the second at 200.

The formula is ln_deaths = -4,06 + 0,91 * ln_confirmed * 1(confirmed < 1) + 0,99* l n_confirmed * 1(1 <= confirmed < 200) + 1,28 * ln_confirmed* 1(confirmed >= 200)

On the graph A7, thanks to the cut off, I can observed that the line is becoming steeper. The deaths is not increasing as fast as the confirmed cases. The adjusted R squared is 0,89. That's a trustful model

Weighted Linear regression : weight with population

This graph represents the Weighted Linear regression : weight with population between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -3,39 + 0,95 * ln_confirmed

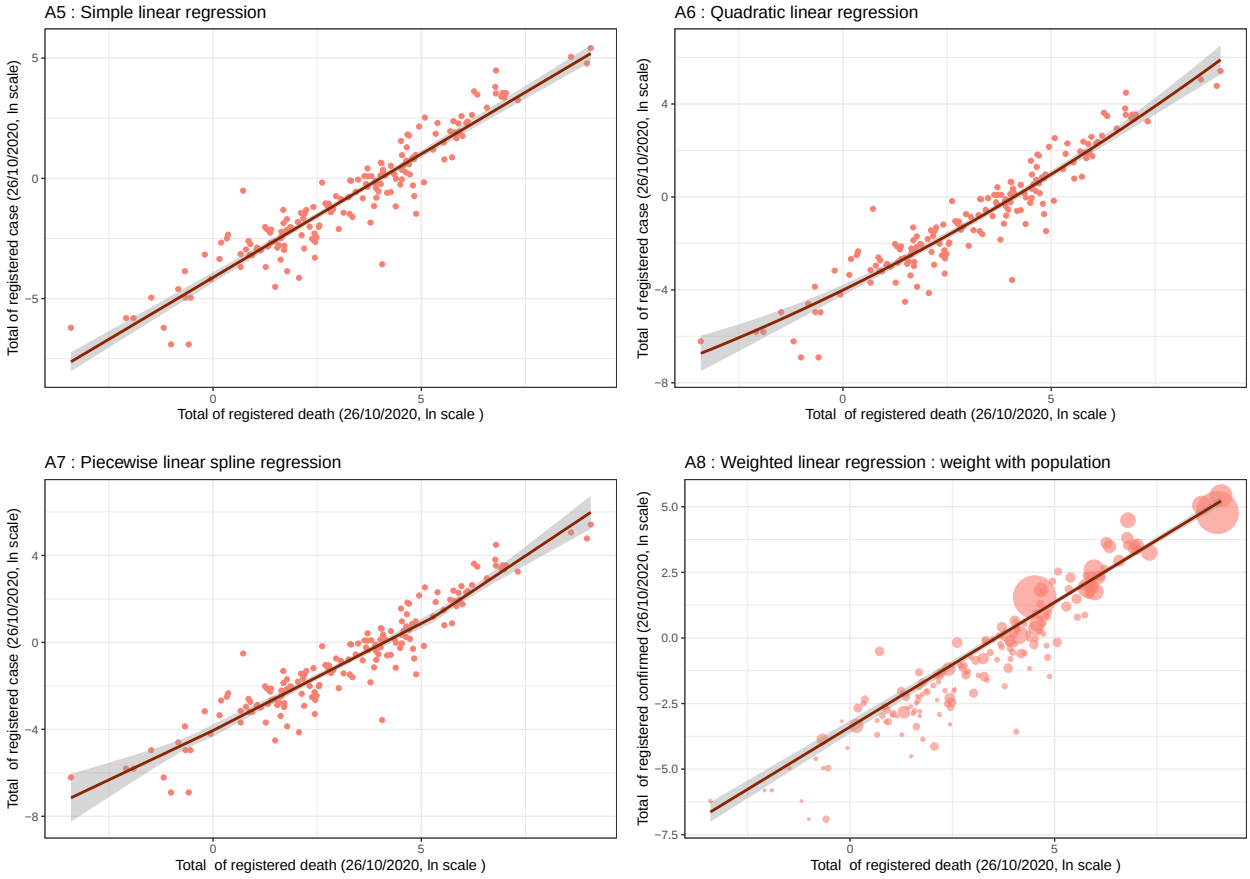
Alpha : -3,39 is the average of ln_death when the confirmed case is one (ln(1)=0).

Beta : the deaths is 9,5 percent higher on average for observation having 10 percent higher case confirmed

In log-log transformation, alpha is usually not meaningfull. The deaths is not increasing as fast as the confirmed cases. I can observe than the countries with the more deaths and confirmed cases are the country with a large population. The adjusted R squared is 0,93. That's a trustful model.

Model Chosen Weighted Linear regression : weight with population

I chose the Weighted Linear regression : weight with population. The adjusted R square of the model is the highest, so it's the model that fits the best. I excluded the PLS and the quadric function, for model complexity reasons. I can see a 0.8 difference between the slope of the linear regression and the slope of the weighted regression, it means that not taking into account the different size of population increase the slope and the predicted deaths. So, I think the pertinent choice is the weight regression



Model summary statistics

	Linear	Quadratic	PLS	weighted linear
(Intercept)	-4.12 ***	-4.01 ***	-4.06 ***	-3.39 ***
	(0.12)	(0.13)	(0.17)	(0.37)
ln_confirmed	1.03 ***	0.88 ***		0.95 ***
	(0.03)	(0.06)		(0.06)
ln_confirmed_sq		0.02 **		
		(0.01)		
lspline(ln_confirmed, cutoff_ln)1			0.91 ***	
			(0.22)	
lspline(ln_confirmed, cutoff_ln)2			0.99 ***	
			(0.05)	
lspline(ln_confirmed, cutoff_ln)3			1.28 ***	
			(0.12)	
nobs	170	170	170	170
r.squared	0.89	0.89	0.89	0.93
adj.r.squared	0.89	0.89	0.89	0.93
statistic	1228.96	722.82	434.66	241.77
p.value	0.00	0.00	0.00	0.00
df.residual	168.00	167.00	166.00	168.00
nobs.1	170.00	170.00	170.00	170.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.